

## Ch 2, Data Analytics: Drawing Conclusions, Part 1

### Integrity of data, p 111- 121

1. **Why is data integrity important? Explain possible effects of using data that lacks integrity.**  
Data integrity is important because it is required to produce useful information, and that it can be accessed. An example of using data that lacks integrity is when searching for records of customers and their purchases, but since the data lacks integrity, some files may be due to poor organisation.

### Timeliness

2. **What does timeliness refer to? Give some examples of data that is not timely.**  
Timeliness refers to its contemporariness and how current it is. An example of data that is not timely is a report on a murder days after the event, when the body and crime scene has had time to change and be altered.

### Authenticity

3. **What are the characteristics of authentic data?**  
Characteristics of authentic data include; that it comes from the author or source, has not been deliberately corrupted, is not faked or disguised as something else, has not been changed without authorisation, is what it claims to be and doesn't misrepresent itself, does not aim to mislead or deceive by pretending to be anything else, and finally, does not lie.
4. **List some challenges to the authenticity of data.**  
Spoofing data (when one or multiple people masquerade as another for fraud, advantage or amusement. Authors may plagiarise other people's publications and steal data, which can be done by editing pictures and documents, making their true source, age and legal status unknown. Legitimate media and advertising productions also add to challenge, as music producers may use auto-tune to repair flat notes that singer have done during recording. CHI in movies and TV is becoming undetectable, as well as photoshopping in magazines. Viral videos such as ad campaigns and fake Youtube videos also grab attention by seeming genuine.
5. **Authenticity techniques. How can you authenticate data, both digital and non-digital?**  
Digital signatures can be used to authenticate digital data, as well as sending a security certificate which is checked by the browser. Checksums can also be used. Finally, email validation can be used to verify that an email address exists and belongs to the correspondent. Non-digital data can be compared to their original documents to compare authenticity with the original author.

### Relevance

6. **What makes data relevant?**  
Data is made relevant based on information that relates to a topic that interests people. It measures how closely a resource corresponds to people's desire for information.

### Accuracy

7. **Distinguish between content and form in terms of accuracy of data.**  
Content refers to correctness and completeness, and revolves around functionality, whereas form refers to the appearance.
8. **Briefly elaborate on the following challenges to data accuracy**

a. **Correctness**

Correctness means that the values stored for a given object are in fact, correct. This is achieved by retrieving truthful results, answers and observations.

b. **Completeness**

Completeness means that data is complete, and acceptable to publish.

c. **Clarity**

Clarity refers to the ease of comprehension towards specific data.

d. **Consistency**

Consistency revolves around keeping the data the same, for example, data may be entered as both "Saint Kilda" and "St Kilda", and while both are correct, they are inconsistent and will produce separate results when queried.

9. Measures to improve accuracy, briefly elaborate on the following measures;

a. **Correctness**

Data quality assurance can be used to cleanse or scrub data, which will identify, remove or repair data that is incomplete, inaccurate, irrelevant or inconsistent. It may also standardise data; for example, by changing all instances of "Street" to "St", or adding data to existing records.

b. **Completeness**

Contacting the original data collectors can assure data completeness, as gaps may have been included in the data, especially that which is unpublished. If the topic is controversial, look at both sides of the argument to see what inconvenient facts may have been ignored. Existence validation checks can also be completed to ensure that essential fields cannot be left empty when data is being entered. Interpolation of data can also be done; to assume data based on trends in order to complete fields and records.

c. **Clarity**

Clarity can be assured by enforcing data formatting and validation rules in spreadsheets and databases that prevent misinterpretation of data. This can be checked by using 'dummy data' to check for ambiguity.

d. **Consistency**

Consistency can be achieved by enforcing consistent data formatting and validation rules. Copy-pasting is far more efficient and accurate than rekeying data, which can create issues such as incorrect data. Data should be standardised so that nothing can be lost due to spelling or wording conventions.